A Probabilistic Framework for Modeling the Variability Across Federated Datasets of Heterogeneous Multi-View Observations

Irene Balelli, Santiago Silva, Marco Lorenzi

A Probabilistic Framework for Modeling the Variability Across Federated Datasets of Heterogeneous Multi-View Observations

Irene Balelli[0000−0002−4593−8217], Santiago Silva[0000−0001−9465−5873], and Marco Lorenzi[0000−0003−0521−2881] for the Alzheimers Disease Neuroimaging Initiative*

Université Côte d'Azur, Inria Sophia Antipolis-Méditeranée, Epione Research Project, France**
{irene.balelli,santiago-smith.silva-rincon,marco.lorenzi}@inria.fr

Abstract. We propose a novel federated learning paradigm to model data variability among heterogeneous clients in multi-centric studies. Our method is expressed through a hierarchical Bayesian latent variable model, where client-specific parameters are assumed to be realization from a global distribution at the master level, which is in turn estimated to account for data bias and variability across clients. We show that our framework can be effectively optimized through expectation maximization over latent master's distribution and clients' parameters. We tested our method on the analysis of multi-modal medical imaging data and clinical scores from distributed clinical datasets of patients affected by Alzheimers disease. We demonstrate that our method is robust when data is distributed either in iid and non-iid manners: it allows to quantify the variability of data, views and centers, while guaranteeing high-quality data reconstruction as compared to the state-of-the-art autoencoding models and federated learning schemes.

Keywords: Federated Learning · Hierarchical Generative Model · Heterogeneity

1 Introduction

The analysis of medical imaging datasets for the study of neurodegenerative diseases, requires the joint modeling of multiple *views*, such as clinical scores and multi-modal medical imaging data. These views are generated through different processes for data acquisition, as for instance Magnetic Resonance Imaging (MRI) or Positron Emission

* Data used in preparation of this article were obtained from the Alzheimers Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu). As such, the investigators within the ADNI contributed to the design and implementation of ADNI and/or provided data but did not participate in analysis or writing of this report. A complete listing of ADNI investigators can be found at: http://adni.loni.usc.edu/wp-content/uploads/how_to_apply/ADNI_Acknowledgement_List.pdf

** This work received financial support by the French government, through the 3IA Côte dAzur Investments in the Future project managed by the National Research Agency (ANR) with the reference number ANR-19-P3IA-0002, and by the ANR JCJC project Fed-BioMed, ref. num. 19-CE45-0006-01. The project was also supported by the Inria Sophia Antipolis - Méediterranée, "NEF" computation cluster.

Tomography (PET). Each view provides a specific information about the pathology, and the joint analysis of all views is necessary to improve diagnosis, for the discovery of pathological relationships or for predicting the disease evolution. Nevertheless, the integration of *multi-views* data, accounting for their mutual interactions and their joint variability, presents a number of challenges.

When dealing with high dimensional and noisy data it is crucial to be able to extract an informative lower dimensional representation to disentangle the relationships among observations, accounting for the intrinsic heterogeneity of the original complex data structure. From a statistical perspective, this implies the estimation of a model of the joint variability across views, or equivalently the development of a joint *generative model*, assuming the existence of a common latent variable generating all views.

Several data assimilation methods based on dimensionality reduction have been developed [4], and successfully applied to a variety of domains. The main goal of these methods is to identify a suitable lower dimensional latent space, where some key characteristics of the original dataset are preserved after projection. The most basic among such methods is Principal Component Analysis (PCA) [7], where data are projected over the axes of maximal variability. More flexible approaches are Auto-Encoders [18], enabling to learn a low-dimensional representation minimizing the reconstruction error.

In some cases, Bayesian counterparts of the original dimensionality reduction methods have been developed, such as Probabilistic Principal Component Analysis (PPCA) [16], based on factor analysis, or, more recently, Variational Auto-Encoders (VAEs) [9]. VAEs are machine learning algorithms based on a generative function which allows probabilistic data reconstruction from the latent space. Encoder and decoder can be flexibly parametrized by neural networks (NNs), and efficiently optimized through Stochastic Gradient Descent (SGD). The added values of a Bayesian formulation is to provide a tool for sampling further observations from the estimated data distribution, and quantify the uncertainty of data and parameters. In addition, Bayesian model selection criteria such as the Watanabe-Akaike Information Criteria (WAIC) [5] allow to perform automatic model selection.

Multi-centric biomedical studies offer a great opportunity to significantly increase the quantity and quality of available data, hence to improve the statistical reliability of their analysis. Nevertheless, in this context, three main data-related challenges should be considered. 1) *Statistical heterogeneity* of *local datasets* (*i.e.* center-specific datasets): observations may be non-identically distributed across centers with respect to some characteristic affecting the output (*e.g.* diagnosis). Additional variability in local datasets can also come from data collection and acquisition bias [8]. 2) *Missing not at random views*: not all views are usually available for each center. 3) *Privacy* concerns: privacy-preserving laws are currently enforced to ensure the protection of personal data (*e.g.* the European General Data Protection Regulation - GDPR[1]), preventing the centralized analysis of data collected in multiple biomedical centers [6,3]. These limitations impose the need for extending data assimilation methods to handle decentralized heterogeneous data and missing views in local datasets.

Federated learning (FL) is an emerging paradigm specifically developed for the decentralized training of machine learning models. In order to guarantee data privacy,

[1] https://gdpr-info.eu/

FL methods are conceived in such a way to avoid any sensitive data transfer among centers: raw data are processed within each center, which only shares local parameters with the master. The standard aggregation method in FL is Federated Averaging (FedAvg) [14], which combines local models via weighted averaging. However, this aggregation scheme is sensitive to statistical heterogeneity, which naturally arises in federated datasets [11], for example when dealing with multi-view data, or when data are not uniformly represented across data centers (e.g. non-iid distributed). In this case a faithful representation of the variability across centers is not guaranteed.

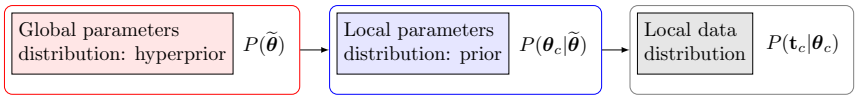


Fig. 1: Hierarchical structure of Fed-mv-PPCA. Global parameters $\widetilde{\boldsymbol{\theta}}$ characterize the distribution of the local $\boldsymbol{\theta}_c$, which parametrize the local data distribution in each center.

We present here the Federated multi-view PPCA (Fed-mv-PPCA), a novel FL framework for data assimilation of heterogeneous multi-view datasets. Our framework is designed to account for the heterogeneity of federated datasets through a fully Bayesian formulation. Fed-mv-PPCA is based on a hierarchical dependency of the model's parameters to handle different sources of variability in the federated dataset (Figure 1). The method is based on a linear generative model, assuming Gaussian latent variables and noise, and allows to account for missing views and observations across datasets. In practice, we assume that there exists an ideal global distribution of each parameter, from which local parameters are generated to account for the local data distribution in each center. The code developed in Python is publicly available at `https://gitlab.inria.fr/ibalelli/fed-mv-ppca`.

The paper is organized as follows: in Section 2 we provide a brief overview of the state-of-the-art and highlight the advancements provided by Fed-mv-PPCA. In Section 3 we describe Fed-mv-PPCA and in Section 4 we show results with applications to synthetic data and to data from the Alzheimers Disease Neuroimaging Initiative dataset (ADNI). Section 5 concludes the paper with a brief discussion.

2 State of the art

The method presented in this paper falls within two main categories: Bayesian methods for data assimilation, and FL methods for heterogeneous datasets. Several methods for dimensionality reduction based on generative models have been developed in the past years, starting from the seminal work of PPCA [16], to Bayesian Canonical Correlation Analysis (CCA) [10], which has been extended to include multiple views and

missing modalities [13], up to more complex methods based on multi-variate association models [15], developed, for example, to integrate multi-modal brain imaging data and high-throughput genomics data. More recent methods for the probabilistic analysis of multi-views datasests include the multi channel Variational Autoencoder (mc-VAE) [1] and Multi-Omics Factor Analysis (MOFA) [2]. MOFA generalizes PPCA for the analysis of multiple-omics data types, supporting different noise models to adapt to continuous, binary and count data, while mc-VAE extends the classic VAE [9] to jointly account for multiple-views data. Additionally, mc-VAE can handle sparse datasets: data reconstruction in testing can be inferred from available views, if some are missing.

Despite the possibility of performing data assimilation and integrate multiple views offered by the above methods, these approaches have not been conceived to handle federated datasets.

Statistical heterogeneity is a key challenge in FL and, more generally, in multi-centric studies [11]. To tackle this problem, Li et al. recently proposed the FedProx algorithm [12], which improves FedAvg by allowing for partial local work (*i.e.* adapting the number of local epochs) and by introducing a proximal term to the local objective function to avoid divergence due to data heterogeneity. Other methods have been developed under the Bayesian non-parametric formalism, such as [19], where the local parameters of NNs are federated depending on neurons similarities.

Despite significant improvements in the handling of statistical heterogeneity have been made since the development of FedAvg, state-of-the-art FL methods are currently essentially formulated for training schemes based on stochastic gradient descent, with principal applications to NNs based models. Beyond the specific application to NNs, we still lack of a consistent Bayesian framework for the estimation of local and global data variability, as part of a global optimization model, while accounting for data heterogeneity. This provides us motivation for the development of Fed-mv-PPCA, a Bayesian framework for data assimilation from heterogeneous multi-views federated datasets.

3 Federated multi-views PPCA

3.1 Problem setup

We consider C independent centers. Each center $c \in \{1, \dots, C\}$ disposes of its private local dataset $T_c = \{\mathbf{t}_{c,n}\}_n$, with $|T_c| = N_c$. We assume that a total of K distinct views have been measured across all centers, and we allow missing views in some local dataset (*i.e.* some local dataset could be incomplete, including only measurements for $K_c < K$ views). For every $k \in \{1, \dots, K\}$, the dimension of the k^{th}-view (*i.e.* the number of features defining the k^{th}-view) is d_k, and we define $d := \sum_{k=1}^{K} d_k$. We denote by $\mathbf{t}_{c,n}^{(k)}$ the raw data of subject n in center c corresponding to the k^{th}-view, hence $\mathbf{t}_{c,n} = \left(\mathbf{t}_{c,n}^{(1)}, \dots, \mathbf{t}_{c,n}^{(K)}\right)$.

3.2 Modeling assumptions

The main assumption at the basis of Fed-mv-PPCA is the existence of a hierarchical structure underlying the data distribution. In particular, we suppose that there exist

global parameters $\widetilde{\boldsymbol{\theta}}$, following a distribution $P(\widetilde{\boldsymbol{\theta}})$, able to describe the global data variability, *i.e.* the ensemble of local datasets. For each center, local parameters $\boldsymbol{\theta}_c$ are generated from $P(\boldsymbol{\theta}_c|\widetilde{\boldsymbol{\theta}})$, to account for the specific variability of the local dataset. Finally, local data $\mathbf{t}_c$ are obtained from their local distribution $P(\mathbf{t}_c|\boldsymbol{\theta}_c)$. Given the federated datasets, Fed-mv-PPCA provides a consistent Bayesian framework to solve the inverse problem and estimate the model's parameters across the entire hierarchy.

We assume that in each center c, the local data corresponding to the k^{th}-view, $\mathbf{t}_{c,n}^{(k)}$, follows the generative model:

$$\mathbf{t}_{c,n}^{(k)} = W_c^{(k)}\mathbf{x}_{c,n} + \boldsymbol{\mu}_c^{(k)} + \boldsymbol{\varepsilon}_c^{(k)}, \tag{1}$$

where $\mathbf{x}_{c,n} \sim \mathcal{N}(0, \mathbb{I}_q)$ is a q-dimensional latent variable, and $q < \min_k(d_k)$ is the dimension of the latent-space. $W_c^{(k)} \in \mathbb{R}^{d_k \times q}$ provides the linear mapping between latent space and observations for the k^{th}-view, $\boldsymbol{\mu}_c^{(k)} \in \mathbb{R}^{d_k}$ is the offset of the data corresponding to view k, and $\boldsymbol{\varepsilon}_c^{(k)} \sim \mathcal{N}\left(0, {\sigma_c^{(k)}}^2 \mathbb{I}_{d_k}\right)$ is a Gaussian noise for the k^{th}-view. This formulation induces a Gaussian distribution over $\mathbf{t}_{c,n}^{(k)}$, implying:

$$\mathbf{t}_{c,n}^{(k)} \sim \mathcal{N}(\boldsymbol{\mu}_c^{(k)}, C_c^{(k)}), \tag{2}$$

where $C_c^{(k)} = W_c^{(k)} {W_c^{(k)}}^T + {\sigma_c^{(k)}}^2 \mathbb{I}_{d_k} \in \mathbb{R}^{d_k \times d_k}$. Finally, a compact formulation for $\mathbf{t}_{c,n}$ (*i.e.* considering all views concatenated) can be derived from Equation (1):

$$\mathbf{t}_{c,n} = W_c\mathbf{x}_{c,n} + \boldsymbol{\mu}_c + \Psi_c, \tag{3}$$

where $W_c, \boldsymbol{\mu}_c$ are obtained by concatenating all $W_c^{(k)}, \boldsymbol{\mu}_c^{(k)}$, and Ψ_c is a block diagonal matrix, where the k^{th}-block is given by $\boldsymbol{\varepsilon}_c^{(k)}$. The local parameters describing the center-specific dataset thus are $\boldsymbol{\theta}_c := \left\{\boldsymbol{\mu}_c^{(k)},\ W_c^{(k)}, {\sigma_c^{(k)}}^2\right\}$. According to our hierarchical formulation, we assume that each local parameter in $\boldsymbol{\theta}_c$ is a realization of a common global prior distribution described by $\widetilde{\boldsymbol{\theta}} := \left\{\widetilde{\boldsymbol{\mu}}^{(k)}, \sigma_{\widetilde{\boldsymbol{\mu}}^{(k)}},\ \widetilde{W}^{(k)}, \sigma_{\widetilde{W}^{(k)}}, \widetilde{\alpha}^{(k)}, \widetilde{\beta}^{(k)}\right\}$. In particular we assume that $\boldsymbol{\mu}_c^{(k)}$ and $W_c^{(k)}$ are normally distributed, while the variance of the Gaussian error, ${\sigma_c^{(k)}}^2$, follows an inverse-gamma distribution. Formally:

$$\boldsymbol{\mu}_c^{(k)}|\widetilde{\boldsymbol{\mu}}^{(k)}, \sigma_{\widetilde{\boldsymbol{\mu}}^{(k)}} \sim \mathcal{N}\left(\widetilde{\boldsymbol{\mu}}^{(k)}, \sigma^2_{\widetilde{\boldsymbol{\mu}}^{(k)}} \mathbb{I}_{d_k}\right), \tag{4}$$

$$W_c^{(k)}|\widetilde{W}^{(k)}, \sigma_{\widetilde{W}^{(k)}} \sim \mathcal{MN}_{k,q}\left(\widetilde{W}^{(k)}, \mathbb{I}_{d_k}, \sigma^2_{\widetilde{W}^{(k)}} \mathbb{I}_q\right), \tag{5}$$

$${\sigma_c^{(k)}}^2|\widetilde{\alpha}^{(k)}, \widetilde{\beta}^{(k)} \sim \text{Inverse-Gamma}(\widetilde{\alpha}^{(k)}, \widetilde{\beta}^{(k)}), \tag{6}$$

where $\mathcal{MN}_{k,q}$ denotes the matrix normal distribution of dimension $d_k \times q$.

3.3 Proposed framework

The assumptions made in Section 3.2 allow to naturally define an optimization scheme based on Expectation Maximization (EM) locally, and on Maximum Likelihood estimation (ML) at the master level (Algorithm 1). Figure 2 shows the graphical model of Fed-mv-PPCA.

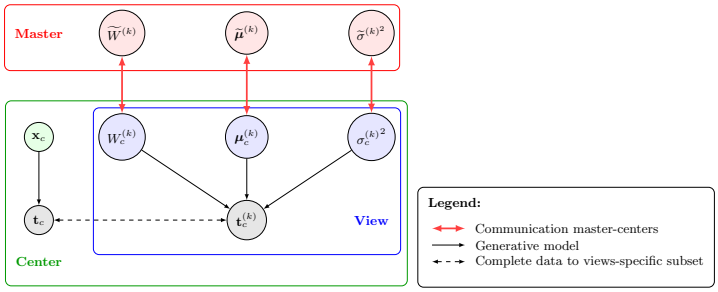


Fig. 2: Graphical model of Fed-mv-PPCA. Thick double-sided red arrows relate nodes which are shared between center and master, while plain black arrows define the relations between the local dataset and the generative model parameters. Grey filled circles correspond to raw data: the dashed double-sided arrow simply highlights the complexity of the dataset, composed by multiple views.

Algorithm 1: Fed-mv-PPCA algorithm

Input : Rounds R; Iterations I; Latent space dimension q
Output: Global parameters $\widetilde{\boldsymbol{\theta}}$

for $r = 1, \dots, R$ **do**
  if $r = 1$ **then**
    Each center c initializes randomly local parameters $\boldsymbol{\theta}_c$;
    I iterations of EM estimation to optimize $\boldsymbol{\theta}_c$;
  else
    Each center initializes $\boldsymbol{\theta}_c$ using $P(\boldsymbol{\theta}_c|\widetilde{\boldsymbol{\theta}})$;
    I iterations of MAP estimation (EM + prior) to optimize $\boldsymbol{\theta}_c$ using $\widetilde{\boldsymbol{\theta}}$ as prior;
  end
  Each center c returns $\boldsymbol{\theta}_c$ to the master;
  The master collects $\boldsymbol{\theta}_c$, $c = 1, \dots, C$ and estimates $\widetilde{\boldsymbol{\theta}}$ through ML;
  The master sends $\widetilde{\boldsymbol{\theta}}$ to all centers
end

With reference to Algorithm 1, the optimization of Fed-mv-PPCA is as follows:

Optimization. The master collects the local parameters $\boldsymbol{\theta}_c$ for $c \in \{1, \dots, C\}$ and estimates the ML updated global parameters characterizing the prior distributions of Equations (4) to (6). Updated global parameters $\widetilde{\boldsymbol{\theta}}$ are returned to each center, and serve as priors to update the MAP estimation of the local parameters $\boldsymbol{\theta}_c$, through the M step on the functional $\mathbf{E}_{p(\mathbf{x}_{c,n}|\mathbf{t}_{c,n})} \ln \left(p(\mathbf{t}_{c,n}, \mathbf{x}_{c,n}|\boldsymbol{\theta}_c) p(\boldsymbol{\theta}_c|\widetilde{\boldsymbol{\theta}}) \right)$, where:

$$p(\mathbf{x}_{c,n}|\mathbf{t}_{c,n}) \sim \mathcal{N}\left(\Sigma_c^{-1} {W_c}^T \Psi_c^{-1} (\mathbf{t}_{c,n} - \boldsymbol{\mu}_c), \Sigma_c^{-1} \right), \Sigma_c := (\mathbb{I}_q + W_c^T \Psi_c^{-1} W_c)$$

and

$$\langle \ln\left(p(\mathbf{t}_{c,n}, \mathbf{x}_{c,n}|\boldsymbol{\theta}_c)\right)\rangle = -\sum_{n=1}^{N_c}\left\{\sum_{k=1}^{K}\left[\frac{d_k}{2}\ln\left({\sigma_c^{(k)}}^2\right) + \frac{1}{2{\sigma_c^{(k)}}^2}\|\mathbf{t}_{c,n}^{(k)} - \boldsymbol{\mu}_c^{(k)}\|^2 + \right.\right.$$
$$\frac{1}{2{\sigma_c^{(k)}}^2} tr\left({W_c^{(k)}}^T W_c^{(k)} \langle \mathbf{x}_{c,n}\mathbf{x}_{c,n}^T\rangle\right)$$
$$\left.\left. -\frac{1}{{\sigma_c^{(k)}}^2}\langle \mathbf{x}_{c,n}\rangle^T {W_c^{(k)}}^T\left(\mathbf{t}_{c,i}^{(k),g} - \boldsymbol{\mu}_c^{(k)}\right)\right] + \frac{1}{2} tr\left(\langle \mathbf{x}_{c,n}\mathbf{x}_{c,n}^T\rangle\right)\right\},$$

Initialization. The latent-space dimension q, the number of local iterations I for EM and the number of communication rounds R (*i.e.* number of complete cycles centers-master) are user-defined parameters. For sake of simplicity, we set here the same number of local iterations for every center. Note that this constraint can be easily adapted to take into account systems heterogeneity among centers, as well as the size of each local dataset. At the first round, each center initializes randomly every local parameter and performs EM through I iterations, maximizing the functional $\langle \ln\left(p(\mathbf{t}_{c,n}, \mathbf{x}_{c,n}|\boldsymbol{\theta}_c)\right)\rangle$.

4 Applications

4.1 Materials

In the preparation of this article we used two datasets.

Synthetic dataset (SD): we generated 400 observations from (1), consisting of $k = 3$ views of dimension $d_1 = 15, d_2 = 8, d_3 = 10$. Each view was generated from a common 5-dimensional latent space. We randomly chose parameters $W^{(k)}, \boldsymbol{\mu}^{(k)}, \sigma^{(k)}$. Finally, to simulate heterogeneity, a randomly chosen sub-sample composed by 250 observations was shifted in the latent space by a randomly generated vector: this allowed to simulate the existence of two distinct groups in the population.

Alzheimer's Disease Neuroimaging Initiative dataset (ADNI)[2]: we consider 311 participants extracted from the ADNI dataset, among cognitively normal (NL) (104 subjects) and patients diagnosed with AD (207 subjects). All participants are associated with multiple data views: cognitive scores including MMSE, CDR-SB, ADAS-Cog-11 and RAVLT (CLINIC), Magnetic resonance imaging (MRI), Fluorodeoxyglucose-PET (FDG) and AV45-Amyloid PET (AV45) images. MRI morphometrical biomarkers were obtained as regional volumes using the cross-sectional pipeline of FreeSurfer v6.0 and the Desikan-Killiany parcellation. Measurements from AV45-PET and FDG-PET were estimated by co-registering each modality to their respective MRI space, normalizing by the cerebellum uptake and by computing regional amyloid load and glucose hypometabolism using PetSurfer pipeline and the same parcellation. Features were corrected beforehand with respect to intra-cranial volume, sex and age using a multivariate

[2] The ADNI project was launched in 2003 as a public-private partnership, led by Principal Investigator Michael W. Weiner, MD. The primary goal of ADNI was to test whether serial magnetic resonance imaging (MRI), positron emission tomography (PET), other biological markers, and clinical and neuropsychological assessments can be combined to measure the progression of early Alzheimer's disease (AD) (see www.adni-info.org for up-to-date information).

linear model. Data dimensions for each view are: $d_{\text{CLINIC}} = 7$, $d_{\text{MRI}} = 41$, $d_{\text{FDG}} = 41$ and $d_{\text{AV45}} = 41$.

4.2 Benchmark

We compare our method to two state-of-the art data assimilation methods: Variational Autoencoder (VAE) [9] and multi-channel VAE (mc-VAE) [1]. In order to obtain the federated version of VAE and mc-VAE we used FedAvg [14], which is specifically conceived for stochastic gradient descent optimization.

4.3 Results

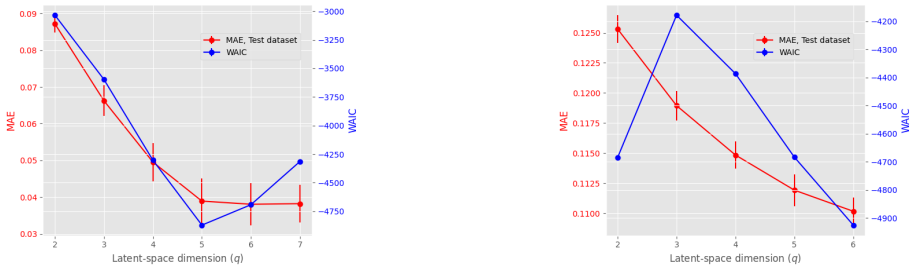
We apply Fed-mv-PPCA to both SD and ADNI datasets, and quantify the quality of reconstruction and identification of the latent space with respect to the increasing number of centers, C, and the increasing data heterogeneity. We investigate also the ability of Fed-mv-PPCA in estimating the data variability and predicting the distribution of missing views. To this end, we consider 4 different scenarios of data distribution across multiple centers, detailed in Table 1.

Table 1: Distribution of Datasets Across Centers.

Scenario	Description
IID	Data are iid distributed across C centers with respect to groups and all subjects dispose of a complete data raw
G	Data are non-iid distributed with respect to groups across C centers: $C/3$ centers includes subjects from both groups; $C/3$ centers only subjects from group 1 (AD in the ADNI case); $C/3$ centers only subjects from group 2 (NL for ADNI). All views have been measured in each center.
K	$C/3$ centers dispose of all observations; in $C/3$ centers the second view (MRI for ADNI) is missing; in $C/3$ centers the third view (FDG for ADNI) is missing. Data are iid distributed across C centers with respect to groups.
G/K	Data are non-iid distributed (scenario G) and there are missing views (scenario K).

Model selection The latent space dimension q is an user defined parameter, with the only constraint $q < \min_k\{d_k\}$. To assess the optimal q, we consider the IID scenario and let q vary. We perform 10 times a 3-fold Cross Validation (3-CV), and split the train dataset across 3 centers. For every test, we perform 100 rounds each consisting of 15 iterations for local optimization. The resulting models are compared using the WAIC criterion. In addition, we consider the Mean Absolute reconstruction Error (MAE) in an hold-out test dataset: the MAE is obtained by evaluating the mean absolute distance between real data and data reconstructed using the global distribution. Figure 3 shows the evolution of WAIC and MAE with respect to the latent space dimension.

Concerning the SD dataset, the WAIC suggests $q = 5$ latent dimensions, hence demonstrating the ability of Fed-mv-PPCA to correctly recover the ground truth latent



(a) SD (b) ADNI

Fig. 3: WAIC score and MAE for (a) the SD dataset and (b) the ADNI dataset. In both figures, the left y-axis scaling describes the MAE while the right y-axis scaling corresponds to the WAIC score.

space dimension used to generate the data. Analogously, the MAE improves drastically up to the dimension $q = 5$, and subsequently stabilizes. For ADNI, the MAE improves for increasing latent space dimensions, and we obtain the best WAIC score for $q = 6$, suggesting that a high-capacity model is preferable to describe this larger dimensional dataset. Despite the agreement of MAE and WAIC for both datasets, the WAIC has the competitive advantage of providing a natural and automatic model selection measure in Bayesian models, which does not require testing data, conversely to MAE.

In the following experiments, we set the latent space dimension $q = 5$ for the SD dataset and $q = 6$ for the ADNI dataset.

Increasing heterogeneity across datasets To test the robustness of Fed-mv-PPCA's results, for each scenario of Table 1, we split the global dataset in C centers, hence we perform 3-CV in each center to obtain local train and test datasets. We compare our method to VAE and mc-VAE. To keep the modeling setup consistent across methods, both auto-encoders were tested by considering linear encoding and decoding mappings [1]. For all methods we consider the MAE in both the train and test datasets, as well as the accuracy score in the Latent Space (LS) discriminating the groups (synthetically defined in SD or corresponding to the clinical diagnosis in ADNI). The classification was performed via Linear Discriminant Analysis (LDA) on the individual projection of test data in the latent space. In what follows we present a detailed description of results corresponding to the ADNI dataset. Results for the SD dataset are in line with what we observe for ADNI, and confirm that our method outperforms VAE and mc-VAE both in reconstruction and in discrimination (not shown for space limitations).

IID distribution. We consider the IID scenario and split the train dataset across 1 to 6 centers. Table 2 shows that results from Fed-mv-PPCA are stable when passing from a centralized to a federated setting, and when considering an increasing number of centers C. We only observe a degradation of the MAE in the train dataset, but this does not affect the performance of Fed-mv-PPCA in reconstructing the test data. Moreover,

Table 2: Heterogeneous Distribution of ADNI Dataset.

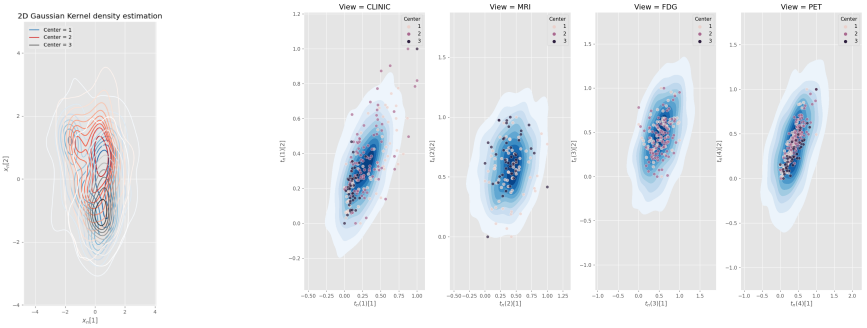
Scenario	Centers	Method	MAE Train	MAE Test	Accuracy in LS
IID	1	**Fed-mv-PPCA**	**0.0804±0.0003**	**0.1113±0.0012**	**0.8839±0.0293**
		VAE	0.1061±0.0277	0.1350±0.0380	0.7364±0.0246
		mc-VAE	0.1392±0.0197	0.1678±0.0279	0.8327±0.0303
	3	**Fed-mv-PPCA**	**0.1059±0.0019**	**0.1102±0.0011**	**0.8962±0.0181**
		VAE	0.1183±0.0355	0.1206±0.0340	0.8681±0.0170
		mc-VAE	0.1606±0.0234	0.1567±0.0216	0.8746±0.0084
	6	**Fed-mv-PPCA**	**0.1258±0.0041**	**0.1116±0.0014**	**0.8930±0.0179**
		VAE	0.1340±0.0433	0.1176±0.0342	0.8071±0.0339
		mc-VAE	0.1837±0.0281	0.1569±0.0200	0.8811±0.0236
G	3	**Fed-mv-PPCA**	**0.1090±0.0041**	**0.1112±0.0013**	**0.8586±0.0272**
		VAE	0.1185±0.0372	0.1194±0.0359	0.7588±0.0581
		mc-VAE	0.1654±0.0262	0.1613±0.0251	0.7985±0.0522
	6	**Fed-mv-PPCA**	**0.1312±0.0113**	**0.1126±0.0014**	**0.8538±0.0354**
		VAE	0.1386±0.0453	0.1188±0.0360	0.7608±0.0492
		mc-VAE	0.1909±0.0323	0.1600±0.0238	0.7872±0.0476
K	3	Fed-mv-PPCA	0.1056±0.0118	0.1300±0.0154	0.8713±0.0227
	6		0.1220±0.0108	0.1287±0.0103	0.8642±0.0162
G/K	3	Fed-mv-PPCA	0.1020±0.0087	0.1301±0.0110	0.7098±0.0329
	6		0.1246±0.0147	0.1313±0.0105	0.7172±0.0315

irrespective of the number of training centers, Fed-mv-PPCA outperforms VAE and mc-VAE both in reconstruction and in preserving subjects separation in the latent space.

Heterogeneous distribution. We simulate an increasing degree of heterogeneity in 3 to 6 local datasets, to further challenge the models in properly recovering the global data. In particular, we consider both a non-iid distribution of subjects across centers, and missing not at random views in some local dataset. It is worth noting that scenarios implying datasets with missing views cannot be handled by VAE nor by mc-VAE, hence in these cases we reported only results obtained with our method.

In Table 2 we report the average MAEs and Accuracy in the latent space for each scenario, obtained over 10 tests for the ADNI dataset. Fed-mv-PPCA is robust despite an increasing degree of heterogeneity in the local datasests. We observe a slight deterioration of the MAE in the test dataset in the more challenging non-iid cases (scenarios K and G/K), while we note a drop of the classification accuracy in the most heterogeneous setup (G/K). Nevertheless, Fed-mv-PPCA demostrates to be more stable and to perform better than VAE and mc-VAE when statistical heterogeneity is introduced.

Figure 4 (a) shows the sampling posterior distribution of the latent variables, while in Figure 4 (b) we plot the predicted global distribution of the corresponding original space against observations, for the G/K scenario and considering 3 training centers. We notice that the variability of centers is well captured, in spite of the heterogeneity of the distribution in the latent space. In particular center 2 and center 3 have two clearly distinct means: this is due to the fact that subjects in these centers belong to two distinct groups (AD in center 2 and NL in center 3). Despite this, Fed-mv-PPCA is able to reconstruct correctly all views, even if 2 views are completely missing



(a) Latent space (b) Original space

Fig. 4: G/K scenario. First two dimensions for (a) sampling from posterior distribution of latent variables $\mathbf{x}_{c,n}$, and (b) predicted distribution $\mathbf{t}_{c,n}^{(k)}$ against real data.

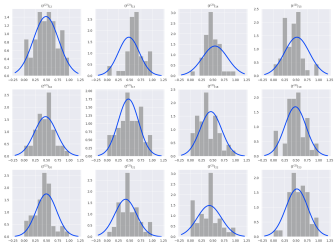


Fig. 5: G/K scenario. Predicted testing distribution (blue curve) of sample features of the missing MRI view against real data (histogram).

in some local datasets (MRI is missing in center 2 and FDG in center 3).

After convergence of Fed-mv-PPCA, each center is supplied with global distributions for each parameter: data corresponding to each view can therefore be simulated, even if some are missing in the local dataset. Considering the same simulation in the challenging G/K scenario as in Figure 4, in Figure 5 we plot the global distribution of some randomly selected features of a missing imaging view in the test center, against ground truth density histogram, from the original data. The global distribution provides an accurate description of the missing MRI view.

5 Conclusions

In spite of the large amount of currently available multi-site biomedical data, we still lack of reliable analysis methods to be applied in multi-centric applications. To tackle this challenge, Fed-mv-PPCA proposes a hierarchical generative model to perform data assimilation of federated heterogeneous multi-view data. The Bayesian approach allows to naturally handle statistical heterogeneity across centers and missing views in local datasets, while providing an interpretable model of data variability. Our applications demonstrate that Fed-mv-PPCA is robust with respect to an increasing degree of heterogeneity across training centers, and provides high-quality data reconstruction, outperforming competitive methods in all scenarios. Further extensions of this work will focus on identifying formal privacy guarantees for Fed-mv-PPCA, to prevent potential private information leakage from the shared statistics, in particular in case of

small local datasets sizes and/or in presence of outliers. To this end, an extension of the proposed framework including Bayesian differential privacy [17] can be foreseen. Other improvements of Fed-mv-PPCA will focus on accounting for sparsity on the reconstruction weights.

References

1. Antelmi, L., Ayache, N., Robert, P., Lorenzi, M.: Sparse multi-channel variational autoencoder for the joint analysis of heterogeneous data (2019)
2. Argelaguet, R., Velten, B., Arnol, D., Dietrich, S., Zenz, T., Marioni, J.C., Buettner, F., Huber, W., Stegle, O.: Multi-omics factor analysis-a framework for unsupervised integration of multi-omics data sets. Mol Syst Biol **14**(6), e8124 (2018)
3. Chassang, G.: The impact of the eu general data protection regulation on scientific research. ecancermedicalscience **11** (2017)
4. Cunningham, J.P., Ghahramani, Z.: Linear dimensionality reduction: Survey, insights, and generalizations. The Journal of Machine Learning Research **16**(1), 2859–2900 (2015)
5. Gelman, A., Hwang, J., Vehtari, A.: Understanding predictive information criteria for bayesian models. Statistics and computing **24**(6), 997–1016 (2014)
6. Iyengar, A., Kundu, A., Pallis, G.: Healthcare informatics and privacy. IEEE Internet Computing **22**(2), 29–31 (2018)
7. Jolliffe, I.T.: Principal components in regression analysis. In: Principal component analysis, pp. 129–155. Springer (1986)
8. Kalter, J., Sweegers, M.G., Verdonck-de Leeuw, I.M., Brug, J., Buffart, L.M.: Development and use of a flexible data harmonization platform to facilitate the harmonization of individual patient data for meta-analyses. BMC research notes **12**(1), 164 (2019)
9. Kingma, D.P., Welling, M.: Stochastic gradient vb and the variational auto-encoder. In: Second International Conference on Learning Representations, ICLR. vol. 19 (2014)
10. Klami, A., Virtanen, S., Kaski, S.: Bayesian canonical correlation analysis. Journal of Machine Learning Research **14**(Apr), 965–1003 (2013)
11. Li, T., Sahu, A.K., Talwalkar, A., Smith, V.: Federated learning: Challenges, methods, and future directions. IEEE Signal Processing Magazine **37**(3), 50–60 (2020)
12. Li, T., Sahu, A.K., Zaheer, M., Sanjabi, M., Talwalkar, A., Smith, V.: Federated optimization in heterogeneous networks. arXiv preprint arXiv:1812.06127 (2018)
13. Matsuura, T., Saito, K., Ushiku, Y., Harada, T.: Generalized bayesian canonical correlation analysis with missing modalities. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 0–0 (2018)
14. McMahan, B., Moore, E., Ramage, D., Hampson, S., y Arcas, B.A.: Communication-efficient learning of deep networks from decentralized data. In: Artificial Intelligence and Statistics. pp. 1273–1282. PMLR (2017)
15. Shen, L., Thompson, P.M.: Brain imaging genomics: Integrated analysis and machine learning. Proceedings of the IEEE **108**(1), 125–162 (2019)
16. Tipping, M.E., Bishop, C.M.: Probabilistic principal component analysis. Journal of the Royal Statistical Society: Series B (Statistical Methodology) **61**(3), 611–622 (1999)
17. Triastcyn, A., Faltings, B.: Federated learning with bayesian differential privacy. In: 2019 IEEE International Conference on Big Data (Big Data). pp. 2587–2596. IEEE (2019)
18. Wang, Y., Yao, H., Zhao, S.: Auto-encoder based dimensionality reduction. Neurocomputing **184**, 232–242 (2016)
19. Yurochkin, M., Agarwal, M., Ghosh, S., Greenewald, K., Hoang, N., Khazaeni, Y.: Probabilistic federated neural matching (2018)